Web Madenciliği Yöntemleri ile Web Loglarının İstatistiksel Analizi ve Saldırı Tespiti

Işıl ÇİNAR, Hasan Şakir BİLGE

Bilgisayar Mühensiliği, Mühendislik Fakültesi, Gazi Üniversitesi, Maltepe, Türkiye
isilscinar@gmail.com, bilge@gazi.edu.tr
(Geliş/Received: 11.03.2015; Kabul/Accepted: 17.02.2016)
DOI: 10.17671/btd.03127

Özet— Webde yer alan bilgilerin doğrusal olmayan bir biçimde hızla artışına paralel olarak web loglarının analiz ihtiyacı da artmıştır. Son yıllarda gerçekleşen siber saldırıların büyük çoğunluğu uygulama katmanında webe yönelik saldırılar olmaktadır. Bu çalışmada veri madenciliği yöntemleri kullanılarak web loglarının 3 aşamada analizi yapılmıştır. Birinci aşamada genel istatistiksel analiz, ikinci aşamada web robotlarının isteklerinin temizlenmesiyle sadece kullanıcılara ait logların analizi ve son aşamada da saldırı tespit edilmesine yönelik analiz yapılmıştır. WEKA yazılımı kullanılarak, web madenciliği teknikleri ile çeşitli çıkarımlarda bulunulmuştur. Saldırı girişimlerinin ve türlerinin tespiti için, WEKA'dan elde edilen sonuçlar ışığında log verisi filtrelenerek açık kaynak web saldırı tespit aracı olan Apache Scalp ile analiz edilmiştir. Web madenciliği ile elde edilen örüntülerden yararlanılarak Apache Scalp ile saldırı girişimleri sayısı %88.7 oranında azalırken, gerçekleştirilen analizin işlem süresi de %90.1 oranında kısalmıştır.

Anahtar Kelimeler— Log analizi, web madenciliği, web güvenliği

Statistical Analysis and Intrusion Detection of Web Logs by Using Web Mining Methods

Abstract— The needs of analysis in web logs have increased in parallel with the information that is increased in a non-linear manner in the web. In recent years, vast majority of the cyber attacks have been against the web in application layer. In this study, web logs have been analyzed in three stages by using the data mining methods. In the first stage, general statistical analysis; in the second stage, analysis of the trace information, which only belongs to the user, with cleaning requests of the web robots, and in the final stage, the analysis for the intrusion detection have been performed. There have been various inferences via the web mining techniques, which perform the data mining approaches to the web data by using WEKA software. In order to detect the intrusion attempts and types, the log data, which is filtered according to the results obtained from WEKA, has been analyzed with Apache Scalp that is open source intrusion detection tool. While the number of the intrusion attempts produced via Apache Scalp by using the patterns obtained from web mining is decreased with 88.7%, the processing time of performed analysis is decreased by the rate of 90.1%.

Keywords— log analysis, web mining, web security

1. GİRİŞ (INTRODUCTION)

İnternet servisleri, web ve mobil uygulamalar günümüzde iletişimle ilgili ihtiyaçlarımızı geniş ölçüde karşılamakta, bununla birlikte muazzam miktarda veri üretilmektedir. Bu verilerin %90'ı önceden tanımlanmış bir yapı ve modelde değildir. Genellikle yapılandırılmamış veri, veri madenciliği ve analiz teknikleri uygulanmadığı sürece kullanışsız olmaktadır [1].

Sistemlerde ağ cihazları logları, güvenlik duvarı logları, saldırı tespit sistemleri logları, web logları gibi birçok çeşitte log tutulmaktadır. Son 10 yılda webde yer alan bilgilerde doğrusal olmayan çok hızlı bir artış söz konusu

olmuştur. Bu büyümeyle webde yer alan bilgilerin analiz ihtiyacı da artmıştır [2]. Bununla birlikte son yıllarda gerçekleşen siber saldırıların yaklaşık %75'inden fazlası uygulama katmanında webe yönelik ataklar olmaktadır [3]. Web sitelerinde etkili bir yönetim ve raporlama için, web sunucular üzerindeki geri beslemelerin alınması gerekmektedir [48].

Web madenciliği veri madenciliğinin alt alanı olup; web madenciliğinde veri madenciliğinde kullanılan yöntemler kullanılmaktadır. Burada web ile ilgili belgelerden ve elde edilen diğer verilerden bilgi ayıklama, analiz etme ve sonuç ortaya çıkarma işlemleri otomatik olarak yapılır [4].

Web madenciliği ile kullanıcılar hakkında detaylı çıkarımlarda bulunulabilmekte, kullanıcıların eğilimlerine göre içerik düzenlenebilmekte, web sitesinin kullanılabilirliğini artırmaya yönelik iyileştirmeler yapılabilmekte ve anomali tespitleri yapılarak çeşitli güvenlik önlemleri alınabilmektedir. Son yıllarda e-ticaret ve çevrimiçi alışveriş hizmetlerinin çoğalmasıyla, bu alanda rekabet sonucu gerçekleştirilen çalışmalar, web madenciliğinin önemini fazlasıyla ortaya çıkarmaktadır [5]. Diğer taraftan kurumsal sayfalarda kurum çalışanlarına yönelik durum değerlendirmesi yapılabilmekte, kullanıcıların karakteristik özellikleri tahmin edilebilmektedir [6].

Web loglarında her bir isteğe karşılık sunucunun verdiği başarılı, başarısız vb. cevaplar durum kodu olarak tutulmaktadır. Literatürdeki bazı çalışmalarda analizler için önişleme aşamasında bu durum kodlarından 400'lü kodlar olarak da bilinen hata kodlarının bulunduğu satırların temizlendiği görülmüştür. Ancak güvenlik açısından analiz gerçekleştirileceği durumlarda özellikle 400 ve 500'lü durum kodlarının incelenmesi gerekmektedir. Çünkü birçok saldırıda saldırgan sayfaya çeşitli istekler göndererek denemelerde bulunmakta ve bu denemeler başarısız olduğunda sunucu çoğu zaman arka arkaya 404 koduyla yanıt vermektedir. Literatürde web saldırı tespitinin gerçekleştirildiği birçok çalışmada bu durumun önişleme aşamasında gözardı edildiği görülmüştür. Bunun sonucunda hem analiz süresinin uzayacağı hem de hatalı sonuçlar alınabileceği düşünülmektedir.

Literatür taramalarında [14-17, 21-23] dikkat çeken bir diğer önemli nokta ise birçok çalışmada web robotları tarafından gerçekleştirilen girdilerin insan girdilerinden ayrılmamış olmasıdır. Ancak çoğu zaman web robotlarının girdileri insan girdileriyle neredeyse eşit oranda olmaktadır ve bu durum analiz sonucunda hatalı çıkarımlarda bulunulmasına sebep olmaktadır.

Bu çalışmada web madenciliği ile veri analizi ve saldırı tespiti için WEKA yazılımı tercih edilmiştir. Kümeleme için Kmeans ve birliktelik kurallarını bulmak amacıyla Apriori algoritmaları kullanılmıştır. WEKA'dan elde edilen sonuçlara göre saldırı türü ve yerini tespit etmek amacıyla Apache Scalp aracı kullanılmıştır. Apache Scalp'ı çalıştırmak ve linux komutlarıyla analizi yapabilmek için linux tabanlı bir güvenlik kontrol işletim sistemi olan BackTrack 5-r3 kullanılmıştır.

Bu çalışmada veri madenciliği yöntemleri kullanılarak web loglarının 3 aşamada analizi yapılmıştır. Birinci aşamada genel istatistiksel analiz, ikinci aşamada web robotlarının isteklerinin temizlenmesiyle sadece kullanıcılara ait logların analizi ve son aşamada da saldırı tespit edilmesine yönelik analiz yapılmıştır.

Bu çalışmanın literatüre katkısı web sunucu loglarının genel istatistiksel analiz, web robotlarının isteklerinin temizlenmesiyle sadece kullanıcılara ait logların analizi ve saldırı tespit edilmesine yönelik analizi gerçekleştirmek üzere önişlemenin 3 aşamada yapılması ve amaca yönelik bu 3 aşamada gerçekleştirilen önişlemelerin farklarının ortaya konulmasıdır. Bu çalışmanın amacı, karmaşık ve düzensiz olan web loglarından web madenciliğinin amacına yönelik olarak gerçekleştirilecek önişleme ile yapılacak analizlerle hem istatistiksel hem de güvenlik açısından aşağıda sunulan çıkarımlarda bulunmaktır:

- Gerçekleştirilen istatistiksel analizler neticesinde kullanıcı eğilimleri ve genel site kullanımı hakkında elde edilen bulgularla site geliştiricilerine web sitesinin iyileştirilmesine ve geliştirilmesine yönelik katkıda bulunmak amaçlanmıştır.
- Web madenciliği yöntemleri ile web saldırı tespitlerini çok daha kısa sürede ve yanlış alarm oranını azaltarak gerçekleştirmek ve elde edilen bulgular neticesinde web sitesinin zafiyetlerini ortaya çıkararak alınabilecek güvenlik önlemleri hakkında çözüm önerileri sunmak hedeflenmiştir.

2. Bölümde, literatür taraması gerçekleştirilerek logların istatistiksel ve güvenlik açısından analizini içeren çalışmalara yer verilmiştir. 3. Bölümde, web saldırı yöntemlerinden bahsedilmiştir. 4. Bölümde genel sistem mimarisi sunularak web madenciliği yöntemleri ile gerçekleştirilen uygulamaya yer verilmiştir. 5. Bölümde Apache Scalp ile log analizine yer verilmiştir. 6. Bölümde ise çalışmanın sonucuna yer verilmiştir.

2. LİTERATÜR İNCELEMESİ (LITERATURE REVIEW)

Log dosyaları bilgisayar sisteminin tarih kitabı olarak kabul edilmektedir [7]. Loglar sistemin çalışmasıyla ilgili önemli bilgiler içermektedir. Bu bilgiler genellikle performans ölçümü, işlemsel profil tespiti, anormalliklerin tespiti, hata ayıklama, güvenlik tehditlerinin tespiti gibi işlemlerde kullanılmaktadır. Son zamanlarda çeşitli veri madenciliği ve makine öğrenme algoritmaları log dosyalarındaki bilgileri analiz etmek üzere yoğun olarak kullanılmaktadır [8].

İnternetteki verilerin sürekli değişmesi, güncellenmesi webden bilgi çıkarımı işlemini zorlaştırmaktadır; web sayfalarının dinamik yapısından dolayı webden bilgi çıkarımı, normal metin tabanlı dokümanlara göre daha zor olmaktadır [54].

Log dosyaları çok fazla bilgi içermektedir, ancak kullanılan dosya formatından bu bilgilerden çıkarım yapmak oldukça zordur. Dolayısıyla verileri analiz etmeye yarayacak bir araca gereksinim duyulmaktadır [9].

Önceleri log analizleri sistem yöneticileri tarafından manuel olarak gerçekleştirilmiştir. Bu durum bazı olayların gözden kaçmasına sebep olabileceği gibi; her geçen gün artan veri miktarı manuel yöntemlerle analizin yetersiz kalmasına sebep olmuş ve analistler çeşitli araçlar geliştirmeye başlamışlardır [10].

Son yıllarda log yönetim sistemleri kurumlar tarafından yaygın olarak kullanılmaya başlanmıştır. Bazı şirketler (IBM, MacAfee ve Splunk gibi) kendilerine özel log yönetim çözümlerini sunmaktadır. Ancak bu sistemlerin uygun donanım gerektirmesi ve log verilerini analiz etmek üzere web kullanım madenciliğini içermemesi problem olmaktadır [11].

Log analiziyle ilgili gerçekleştirilen literatür çalışmalarının büyük bir kısmında web logları analiz edilmiştir. Bu çalışmada doğrudan web logları üzerinde analiz gerçekleştirilmiş ve literatür taramalarına bu yönde ağırlık verilmiştir.

2.1. Logların İstatistiksel Analizi ile İlgili Yapılan Çalışmalar (Studies Concerned with Statistical Analysis of Logs)

Web sayfalarına giriş isteklerinin bilinmesi, sayfalara erişim sıklığının hesaplanması, sayfaların birlikte ziyaret edildiği diğer sayfaların tespit edilmesi gibi istatistiksel yaklaşımlarla kullanıcı davranışlarını öğrenmeyi ve bu sayede site ile ilgili düzenlemeler yapmayı sağlamaktadır [12].

Web madenciliği yaklaşımında yaygın olarak ilk önce web log dosyaları önişleme aşamasından geçirilir. Bunun en önemli nedeni web log dosyalarında çok miktarda gereksiz bilginin olmasıdır. Önişleme sayesinde web log dosyalarının yönetilmesinin daha kolay bir şekilde yapılmasına olanak sağlanmaktadır.

Özakar ve Püskülcü (2002) çalışmalarında web madenciliği yöntemlerinin entegrasyonu ile oluşmuş bir veritabanından yararlanmışlardır. Burada İzmir Yüksek Teknoloji Enstitüsü'nün sayfalarından web madenciliği ile sistem mimarisi çıkarılmıştır. Bu çalışmanın farklı bir özelliği analiz işlemi yapmak yerine sistemin mimarisinin tasarlanması ve bu mimarinin geliştirilmesidir [14].İsmail Haberal (2007), Başkent Üniversitesi web sitesinin geliştirilmesi amacıyla web günlük erişimlerini veri içindeki sınıflandırma bilgisini veren java platformunda geliştirilmiş WEKULA yazılımı ile gerçekleştirerek; serbest kullanıma sunulmuş olan WUM, WUMPREP ve WEKA yazılımları ile veri önişlemesini gerçekleştirmiştir [12].

Daş (2008) yaptığı çalışmada yol analizi yöntemi ile web kullanıcı erişim kütük dosyalarından bilgi çıkarımı yapmış ve birliktelik kuralları yöntemi ile web sayfaları arasındaki ilişkileri belirlemiştir. İstatistiksel analiz yöntemi ile Nihuo ve Web log analiz programlarını kullanarak web sitesinin genel kullanımına ilişkin detaylı bilgiler elde etmiştir [6]. Hussain, Asghar ve Masood (2010) çalışmalarında web log dosyalarını; client log dosyası, proxy log dosyası ve sunucu log dosyası olarak ele almışlardır. Çalışmada log dosyalarında yapılan veri madenciliği işlemlerinin %80'i ön işleme aşamasına ait çıkmıştır. Bu oldukça yüksek bir orandır ve bu aşamanın önemini göstermektedir. Bu aşama sonraki aşamaların yani örüntü analizi ve örüntü çıkarma gibi işlemlerin çok etkili ve kolayca yapılmasına imkan sağlar. Örüntü çıkarmaya yönelik en yaygın kullanılan özellikler web log dosyalarındaki zaman damgası, IP adresi, URL adresi ve User-agent bilgisidir [13].

Mustafa Turan (2011), web madenciliği teknikleriyle makine öğrenme tekniklerini birlikte kullanarak hibrit bir yapı geliştirmiştir. Web sayfalarındaki verileri yapısal olarak incelemenin yanında metin analiz işlevlerini de gerçekleştirmiştir [18].

Güncel Sarıman (2011), mevcut yazılımlarla gerçekleştirilen web erişim kütük analizinin daha kısa sürede geçekleştirilebilmesi için paralel programlama teknikleri kullanmıştır. Paralel ve seri algoritmalar çalıştırılarak analizler süre bakımından karşılaştırılmıştır [19].

Özseven ve Düğenci (2011), web sitesi erişim kayıtlarının daha kolay analiz edilmesini sağlamak için log dosyalarını temizleyerek veritabanına aktaran LOG PreProcessing isminde bir yazılım hazırlamışlardır. Web kullanım madenciliğinin en önemli ve uzun süren aşaması olan ön işlem süreci bu yazılım yardımıyla gerçekleştirildikten sonra standart SQL ifadeleri yardımıyla da siteye ait istatistiki bilgiler elde edilebileceği belirtilmiştir [47].

Web kullanım madenciliği, web madenciliği konusunda önemli ve hızla gelişen; birçok araştırmanın yapıldığı bir alandır [49]. Sisodia ve Verma (2012), web kullanım madenciliği ile ilgili çalışmışlar ve 1 ay süreyle NITR (National Institute of Technology Raipur) web sunucu loglarını kaydedip üzerinde bir yazılım aracılığıyla analiz yapmışlardır [15]. Veri madenciliği yöntemleri ile erişim loglarını analiz edip çeşitli web kullanım erişim modelleri ortaya çıkarmışlardır. Burada istatistiksel tanımlamalar ve birliktelik kuralları kullanmışlardır. Nihayetinde kullanıcıların band genişliği kullanımı, erişim modelleri ve haftalık ziyaretçi sayısı gibi bilgiler elde edilmiştir. Bu tür çalışmalarda, en zor konu; insanların talepleri

sonucunda oluşan girdiler ile web robotlarının talepleri sonucunda oluşan girdiler arasındaki ayrımın iyi yapılmasıdır [15].

Web kullanım madenciliği ile terör saldırıları takibi gerçekleştirilen çalışmada [17] doğru ve kaliteli analiz yapabilmek için web madenciliğinin ön işleme adımı kullanılmıştır. İlgisiz verilerin temizlenmesi ve veri azaltma işlemi iki açıdan ele alınmıştır. İlki birçok incelenen çalışmada olduğu gibi .gif, .jpeg, .css vb. URL isteklerinin temizlenmesi adımıdır. İkinci temizleme işlemi ise hata durum kodları ele alınarak gerçekleştirilmiştir. Hatalı isteklerin madencilik işinde çok kullanışsız olduğu, durum kodlarının kontrol edilerek bu isteklerin temizlenmesi gerektiği belirtilmiştir [17]. Hatalı kodların tespiti için 400 ve 500 durum kodları arasında yer alan isteklerin belirlenmesi gerekmektedir [18].

Kadir Can Burçak (2012), Kırıkkale Üniversitesi web sunucularına ait kullanıcı erişim günlüklerinde önişleme gerçekleştirerek log analizini Nihuo Web Log Analiz Programı ile gerçekleştirmiş ve site kullanımına ait istatistiksel sonuçlar elde etmiştir [9].

2.2. Logların Güvenlik Açısından Analizi ile İlgili Yapılan Çalışmalar (Studies Concerned with the Analysis of Logs with Concern of the Security)

Literatür taramalarında loglar üzerinde analizler gerçekleştirilirken özellikle web saldırılarına yönelik çalışmaların yapıldığı görülmüştür.

Bilişim sistemlerine yönelik saldırılar aktif ve pasif olmak üzere temelde iki yöntemle belirlenmektedir. Aktif saldırı belirleme ve engelleme sistemleri genellikle ağ/host tabanlı çalışmakta olup; anlık ağ trafiği ya da işletim sistemi fonksiyonlarını kullanarak saldırıları belirlemekte ve engellemektedir [20]. Pasif saldırı belirleme sisteminde ise loglar incelenmektedir. Saldırıların büyük bir kısmı loglardaki anormallikler incelenerek belirlenebilmektedir [20].

Vigna, G. ve diğerleri (2003) web uygulamalarındaki karmaşıklığın artmasının web saldırılarına da zemin hazırladığını açıklamaktadırlar. İstemcilerden gelen isteklerin görüntülenmesi ve analiz edilmesine yönelik web saldırı tespit aracı geliştirmişlerdir. Bu aracı güvenlik mimarisi ile bütünleştirmişlerdir [21].

Auxilia ve Tamilselvan (2010) ise makalelerinde önemli web ataklarını; "Sofistike HTTP Saldırıları", "SQL Enjeksiyon Saldırıları" ve "XSS" başlıkları altında incelemişler ve negatif güvenlik modeli geliştirmişlerdir [22].

Ağ saldırısı tespiti için yapılan diğer bir çalışmada bir üniversitedeki yerel alan ağından elde edilen loglardan yararlanılmıştır. Log dosyaları çok büyük olduğu için sapan veriler tespit edilerek anormallikler yakalanmaya çalışılmıştır. Ağ saldırıları "servis kaynaklarını tüketmeye yönelik saldırılar" ve "zararlı çeşitli illegal bilgilerin elektronik posta sunucuları aracılığı ile gönderilmesi ile önemli bilgilere zarar verilmesi" başlıkları altında ele alınmıştır [23].

Elmas Yıldız (2010) çalışmasında DoS ataklarından biri olan Brute Force tipi bir saldırının veri madenciliği teknikleriyle tespit edilip engellendiği bir saldırı tespit uygulaması tasarlamıştır [24].

Salama, Marie, El-Fargary ve Helmy (2011) web saldırı tespiti için web sunucu logları önişlemesi ile ilgili yaptıkları çalışmada web saldırı ve web kullanım madenciliği için yapılan önişlemelerin farklarını açıklayarak saldırı tespitinde önişlemenin öneminden bahsetmişlerdir [25].

Patil (2012) web loglarının önişlenmesiyle kötüye kullanım ve anomali tabanlı saldırı tespiti üzerine bir çalışma gerçekleştirmiştir. Apriori ile birliktelik kuralları oluşturarak normal olmayan davranışları select, insert, <script> vb. SQL ve XSS saldırısına işaret eden anahtar kelimeleri kullanarak tespit etmiştir [26].

Gržinic, Kišasondi ve Šaban (2013), Apache web sunucu loglarının önişleme aşamasında sadece 400 ve 500 arasındaki durum kodlarının yer aldığı satırları alarak web madenciliği işlemleri için WEKA programında yer alan çeşitli sınıflandırma algoritmalarıyla anomali sınıflandırması gerçekleştirmişlerdir [17].

Mabzool ve Lighvan (2014) web madenciliği teknolojisini kullanarak, çevrimdışı saldırı tespit sistemini gerçekleştirmek üzere K-means kümeleme algoritmasını kullanarak anomali davranışları tespit etmişlerdir [27].

3. WEB SALDIRI YÖNTEMLERİ (METHODS OF WEB ATTACKS)

Log analizi gerçekleştirilirken istatistiksel sonuçlar elde etmek kullanıcılara geri dönüş, web sayfasının etkin kullanımı, sayfa yapısı vb. açısından çok önemli olduğu gibi saldırılara yönelik analizlerin gerçekleştirilmesi de büyük önem taşımaktadır. Bu konuyla ilgili birçok ticari, açık kaynak araçlar mevcut olsa da çoğu zaman bu araçlar yanlış alarm üretmekte ve kullanıcı gözüyle analiz gerekmektedir.

Open Web Application Security Project (OWASP), 2007, 2010 ve 2013 yıllarında "OWASP TOP 10" başlığı altında web uygulamaları güvenliğine yönelik kritik riskleri açıklamıştır [28,29]. İncelenen birçok çalışmada OWASP'ın hazırlamış olduğu bu kritik riskler temel alınmıştır.

2013 yılında, OWASP, XSS ve SQL Enjeksiyon açıklıklarını web tabanlı sistemlerdeki en ciddi açıklıklar

olarak raporlamıştır [29,30]. Bu ataklar çok güçlü olmalarının yanı sıra kolay uygulanabilen saldırı tekniklerdir. Web uygulamalarına yönelik gerçekleştirilen bu saldırıların giderilmesine yönelik yapılan çalışmalarda karşılaşılan en önemli problemlerden biri de yanlış pozitif ve yanlış negatif alarmlarla sıkça karşılaşılmasıdır [31]. OWASP'ın açıkladığı en önemli 10 güvenlik riski Tablo 1'de sunulmuştur [32].

Tablo 1. OWASP 10 güvenlik riski
(OWASP 10 Security Risk)

OWASP 10 Güvenlik Riski
SQL Enjeksiyon
Siteler Arası Betik Yazma (XSS)
İhlal edilmiş kimlik doğrulama ve oturum yönetimi
Emniyetsiz doğrudan nesne referansı
Güvenlik ayarları hataları
Hassas veri teşhiri
İşlev seviyesi erişim kontrol eksikliği
Siteler Ötesi İstek Sahteciliği (CSRF)
Bilinen güvenlik açıklıkları ile bileşenleri kullanma
Geçersiz yönlendirme ve iletimler

Bu listenin başında bulunan ve en sık karşılaşılan ataklardan biri olan SQL Enjeksiyon, kullanıcı girdileri ile uygulamaları istismar etmek amacıyla veritabanında SQL komutları çalıştırma tekniğidir [51]. Açıklık kullanıcı girdileriyle SQL deyimini çalıştıracak izin verilmemesi gereken özel karakterlerin kullanılmasıyla ortaya çıkmaktadır [52].

Diğer bir önemli atak olan XSS ise istemci tabanlı kodun, HTML kodlarının arasına eklenerek, kurbanın tarayıcısında zararlı kodun çalıştırılabilmesi olarak tanımlanmaktadır [50]. XSS halen web uygulamaları için büyük bir problem olmaya devam etmektedir. XSS saldırılarını azaltmak için kaynak kodun detaylı incelenmesine ihtiyaç duyulmaktadır [53].

4. WEB SUNUCUSU LOGLARI ÜZERİNDE YAPILAN ANALİZ ÇALIŞMASI (THE ANALYSIS STUDY ON THE WEB SERVER LOGS)

Web log dosyaları web sitelerinin kullanımıyla ilgili verileri depolamaktadır. Bu verilerin analizi son derece önemlidir. Genellikle web log dosyaları farklı formatlarda kaydedilmekte bu durum analiz edilmelerini zorlaştırmaktadır [33].

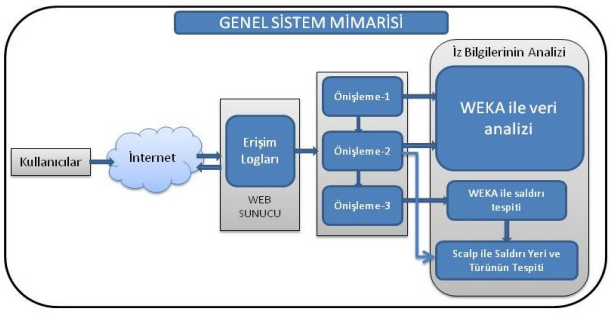
Web log dosyalarında 2 farklı format görülmektedir. Bunlardan biri "Common Log Format" ve diğeri ise "Combined Log Format"tır. Bu makalede çalışılan erişim dosyasındaki kayıt formatı ikinci türdendir. Aşağıda bu formata uygun web log erişim dosyasının alanları sıralanmıştır:

- İsteği gerçekleştiren IP numarası,
- İsteğin yapıldığı tarih ve saat,
- İstek yapılan URL,
- Sunucunun verdiği yanıt olan durum kodu,
- Gönderilen dosyasının boyutu,
- İsteğe nereden ulaşıldığı,
- İstekte bulunan kullanıcıya ait tarayıcı bilgisi

Çalışmada bir yemek web sitesine ait web erişim log dosyası 3 farklı yöntemle analiz edilmiştir.

1. WEKA yazılımı kullanılarak web erişim loglarından istatistiksel analizler yapılmıştır.
2. WEKA yazılımı kullanılarak saldırı tespitine yönelik çeşitli çıkarımlarda bulunulmuştur.
3. WEKA'dan elde edilen sonuçlar ışığında filtrelenmiş veri kullanılarak web saldırı tespitinde kullanılan Apache Scalp aracında saldırı ve açıklık tespiti gerçekleştirilmiştir.

Çalışmanın genel sistem mimarisi Şekil 1'de sunulmuştur.



Şekil 1. Genel sistem mimarisi (Overall system architecture)

4.1. Önişleme Adımları (Preprocessing Steps)

Veri analizinde yapılacak çalışmanın kalitesi bu aşamayla doğrudan ilişkilidir [34]. Bu aşama web kullanım madenciliğinin en önemli aşaması olarak kabul edilmektedir; çünkü etkili bir şekilde yapıldığında zaman ve kaynak tasarrufu sağlamaktadır [35].

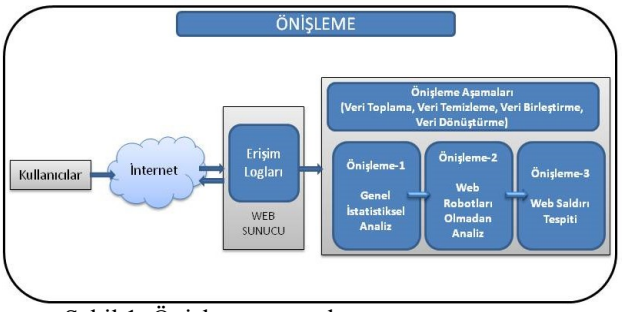
Veri önişleme web madenciliğinde amaca yönelik olarak değişkenlik göstermektedir. Log dosyasında stil bilgileri ve grafik bilgileri gibi önemli olmayan nesneler atılarak veri temizlemesi yapılmıştır [36]. Ancak eğer resim ağırlıklı siteler ve haber siteleri söz konusu olursa bu bilgilerin atılması doğru olmayabilir [37].

Web kullanım madenciliğinde özellikle 200 serili durum kodlarının yer aldığı isteklerin analiz edilmesi gerekmektedir. Web loglarından saldırı analizi yapılacağında ise web kullanım madenciliği önişlemesinin aksine anomali tespiti içerebileceğinden 400 serili durum kodlarının korunması gerekmektedir. 200 serili isteğin başarılı olduğunu gösteren durum kodlarının yer aldığı isteklerin ise silinmesi

gerekmektedir. Ancak saldırı başarılı olduğunda hata kodlarının ardından başarı durum kodu alınacağı dikkate alınmalıdır [25].

Önişlemenin etkin yapılması analizin hızını, güvenilirliğini, genel olarak performansını ciddi oranda etkilemektedir. Veri madenciliği yöntemleri ile web saldırı tespiti yapan çalışmaların [38, 39, 40] birçoğunda önişleme aşamasında hata durum kodları veya başarı durum kodlarına yönelik önişleme çalışması yapılmadığı görülmüştür.

Bu çalışmada web madenciliği amacına yönelik olarak önişleme 3 aşamada gerçekleştirilmiş; her bir aşamaya özel önişleme dosyası oluşturularak web madenciliği analizi yapılmıştır. Şekil 2'de bu aşamalar gösterilmektedir.



Şekil 1. Önişleme aşamaları (Preprocessing steps)

Her bir aşamaya özel gerçekleştirilen önişleme bilgileri aşağıda sunulmuştur.

1. Web kullanım madenciliğine yönelik genel istatistiksel sonuçlar çıkarma amacıyla gerçekleştirilen önişleme adımları aşağıda sunulmuştur:

   - Kimlik ve kullanıcı alanlarında veri tutulmadığı için bu alanlar silinmiştir.
   - Resim ve müzik dosyaları silinmiştir.
   - WEKA'da yapılacak örüntü keşfi aşamaları için tarih formatı "gün/ay/yıl" formatına dönüştürülmüştür.
   - Metot, protokol ve URL alanları birleştirilmiştir.

2. İlk maddede gerçekleştirilen önişlemede web robotları erişimleri de bulunmaktadır. Web robotları (crawler, spider, bot vb.) "robots.txt" dosyasına erişmekte ve yöneticinin verdiği izinlere göre işlem yapmaktadırlar. İz bilgilerinin kaydadeğer bir bölümünü web robotları oluşturduğundan bu isteklerin temizlenmesi analizde kullanıcı odaklı sonuç çıkarma açısından çok önemlidir.
3. Web saldırı tespitine yönelik analiz yapmak amacıyla ilk iki maddede gerçekleştirilen önişlemeye ilave olarak 400 serili hata durum kodlarının dışındaki durum kodlarının yer aldığı istekler silinmiştir.

4.2. Weka ile Örüntü Keşfi ve Analizi (Pattern Discovery and Analysis by using Weka)

WEKA makine öğrenme ve veri madenciliği için kapsamlı bir araçtır [41]. WEKA, açık kaynak kodlu ve java platformu üzerinde geliştirilmiş bir programdır [42]. WEKA'ya verilerin yüklenmesi için çeşitli yöntemler bulunmaktadır:

- Verileri dosyadan açma
- Verileri URL'den açma
- Verileri veritabanından açma

Bu çalışmada veriler veritabanı aracılığı ile WEKA'ya aktarılmıştır.

4.2.1. Önişleme-1 verisi ile analiz (Analysis with preprocessing-1 data)

WEKA ile örüntü keşfi ve analiz için önişleme-1 adımından elde edilen veri tablosu, veritabanı bağlantısıyla yüklenmiştir. WEKA'ya veriler yüklendikten sonra WEKA içerisinde yer alan önişlem adımında görüntülenmektedir. Bu aşamada filtreleme işlemleri ve istatistiksel analizler yapılabilmektedir. Erişim log dosyasının analizi sonucunda ortaya çıkan istatistiksel rakamlar Tablo 2'de listelenmiştir.

Tablo 2. Erişim log dosyası istatistiksel sonuçlar
(Access log file statistical results)

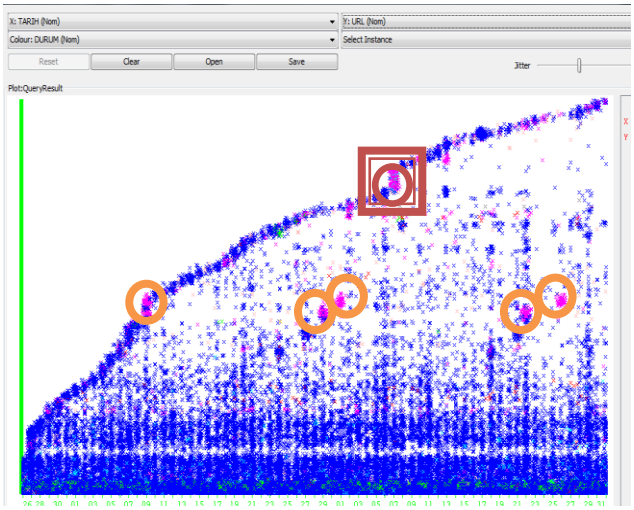
Erişimler (Önişleme-1 için)	
Toplam Veri	685.272
Önişlem Sonucunda Veri Sayısı	64.427
Gün Boyuca Ortalama Erişim	961.59
Ziyaretçi Başına Ortalama Erişim	10.87
Başarılı Olan İstek Sayısı (200 serili kod)	57.561
Hatalı İstek Sayısı (400 serili kod)	1393
Ziyaretçi	
Toplam Ziyaretçi	5925
Ortalama Günlük Ziyaretçi	88.43

WEKA'nın Visualize sekmesinden analiz edilen verinin istenilen alanları arasındaki ilişki yayılım grafiği görülebilmektedir. Bu sekmede alanların dağılımlarına göre kesişim noktaları görülebildiğinden kümeleme analizine olanak sağlamaktadır [43].

İncelenen ana log verisi için Tarih, URL ve durum alanlarına göre işlem yapıldığında Şekil 3'deki sonuç elde edilmiştir. Burada X ekseni tarih alanını, Y ekseni URL alanını, ekranın en altında görüldüğü gibi farklı renkler de durum alanını temsil etmektedir. Bu 3 alana göre pembe renk ile temsil edilen 404 hata durum kodu için Şekil 3'de işaretlenen 6 yerde küme oluştuğu görülmektedir.

Visualize ekranında koordinat düzleminde yer alan her bir kaydın üstüne tıklandığında detaylı bilgi görüntülenmektedir. Bu kümelerin bulunduğu alanlar bu şekilde detaylı incelendiğinde çoğu kümenin arama robotlarının kısa sürede arka arkaya gerçekleştirdiği indeks kayıtlarından oluştuğu görülmektedir. Şekil 4'de tarayıcı alanları incelendiğinde isteklerin web robotlarından geldiği görülmektedir.

Sadece kırmızı renkle işaretlenen ve kare ile gösterilen küme detaylı incelendiğinde robot olmadığı anlaşılan bir IP numarasından, çok kısa süre içerisinde fazla sayıda istekte bulunulduğu ve istek yapılan URL alanı incelendiğinde saldırı denemelerinde bulunulduğu tespit edilmiştir. Şekil 5'de görüldüğü gibi istek herhangi bir web robotundan gelmemektedir. Arka arkaya aynı IP 'den gelmektedir. İstek yapılan URL incelendiğinde saldırı girişiminde bulunulduğu görülmektedir.



Şekil 3. WEKA'nın Visualize sekmesinde gerçekleştirilen analiz (The analysis carried out at the visualize tab in WEKA)

```
Instance:  56732
     IPNO: 217.69.133.67
    TARIH: 22/Oct/2012
      URL: "GET /tr/10928-mehtap-cafe?format=raw&restoranId=10928&view=restoran_resimler HTTP/1.0"
    DURUM: 404
    BOYUT: 1370.0
BASVURULAN: "-"
 TARAYICI: "Mozilla/5.0 (compatible; Mail.RU_Bot/2.0)"

Plot : Master Plot
Instance:  56738
     IPNO: 217.69.133.67
    TARIH: 22/Oct/2012
      URL: "GET /tr/10890-vera-restaurant?format=raw&restoranId=10890&view=restoran_resimler HTTP/1.0"
    DURUM: 404
    BOYUT: 1370.0
BASVURULAN: "-"
 TARAYICI: "Mozilla/5.0 (compatible; Mail.RU_Bot/2.0)"

Plot : Master Plot
Instance:  56741
     IPNO: 217.69.133.67
    TARIH: 22/Oct/2012
      URL: "GET /tr/10887-afrodit-restaurant?format=raw&restoranId=10887&view=restoran_resimler HTTP/1.0"
    DURUM: 404
    BOYUT: 1370.0
BASVURULAN: "-"
 TARAYICI: "Mozilla/5.0 (compatible; Mail.RU_Bot/2.0)"
```

Şekil 4. Analiz ile ilgili detaylı bilgi (Detailed information about Analysis)

```
Instance:  41754
     IPNO: 65.111.180.205
    TARIH: 07/Oct/2012
      URL: "GET //WEB-INF/web.xml HTTP/1.1"
    DURUM: 404
    BOYUT: 1175.0
BASVURULAN: "-"
 TARAYICI: "Mozilla/5.0 (compatible; MSIE 9.0;

Plot : Master Plot
Instance:  41756
     IPNO: 65.111.180.205
    TARIH: 07/Oct/2012
      URL: "GET /db/main.php HTTP/1.1"
    DURUM: 404
    BOYUT: 1175.0
BASVURULAN: "-"
 TARAYICI: "Mozilla/5.0 (compatible; MSIE 9.0;

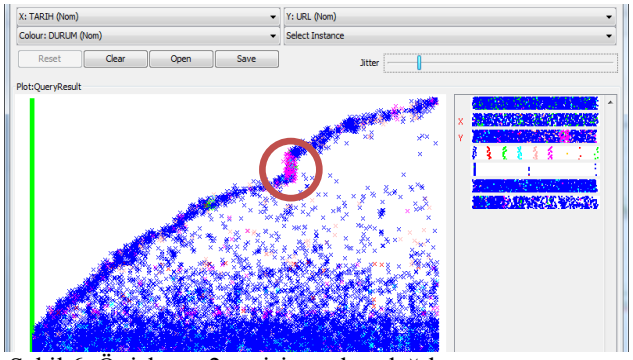
Plot : Master Plot
Instance:  41757
     IPNO: 65.111.180.205
    TARIH: 07/Oct/2012
      URL: "GET /?page=../../../../../../../../../etc/passwd%00.jpg HTTP/1.1"
    DURUM: 302
```

Şekil 5. Saldırı girişimleri içeren istekler (Requests with attack attempts)

4.2.2. Önişleme-2 verisine göre analiz (Analysis of data by the preprocessing-2)

Önişleme-1 verisine kümeleme ve birliktelik algoritmaları uygulandığında web robotlarının kayıtlarının çok fazla olmasından dolayı; çıkan sonuçlar kullanıcı, saldırı vb. analizlerinde yanılmalara sebep olabilmektedir. Web robotlarının yer aldığı log satırları temizlenerek elde edilen Önişleme-2 verisine kümeleme ve birliktelik algoritmaları uygulanarak anlamlı sonuçlar elde edilmiştir.

WEKA'nın Vizualize sekmesinden tarih, URL ve durum alanları için yayılım dağılımına bakıldığında Şekil 6'da görüldüğü gibi sadece 1 küme oluştuğu ve bu kümenin de daha önce önişleme-1 verisinde saldırı girişiminde bulunulduğu tespit edilen kümeyle aynı küme olduğu görülmektedir.



Şekil 6. Önişleme-2 verisi yayılım dağılımı (Preprocessing-2 data spread distribution)

Böylece Önişleme-2 verisinde web robotlarının temizlenmesiyle daha etkin sonuç elde edilmiştir. Yine aynı veriye kümeleme algoritmalarından Kmeans; URL, tarih ve durum alanları için uygulanmıştır. Şekil 7'de elde edilen sonuç sunulmuştur.

```
                                                Cluster#
                                                       0
                                                 (24400)
=========================================================
                                               04/Sep/2012
POST /iphone/restoran_bilgi.php HTTP/1.1"
                                                     200

                                                       1
                                                  (4820)
=========================================================================
                                               29/Aug/2012
"GET /index.php?option=com_egbcaptcha&width=85&height=35&characters=4 HTTP/1.1"
                                                     200
```

Şekil 7. K-means kümeleme algoritması ile yapılan analiz
(Analysis by K-means clustering algorithm)

Burada başarılı (200 durum kodu) 2 istek için küme oluştuğu görülmektedir ve istek yapılan sayfaların da en çok ziyaret edilen sayfalardan oldukları bilinmektedir. 404 hata durum koduyla ilgili Visualize sekmesinde yayılım dağılımında küme oluştuğu halde burada K-means kümeleme algoritmasıyla oluşmamıştır. Çünkü web loglarında genel olarak, özellikle geniş tarih aralığında çalışma yapıldığı durumlarda, 200 serili durum kodlarının yer aldığı istek sayısına göre 400 serili hata durum kodlarının yer aldığı istek sayısı oldukça azdır. Dolayısıyla özellikle saldırı tespitine yönelik analiz yapılacağında 400 serili hata durum kodları haricindeki kayıtların önişleme aşamasında silinmesi gerekmektedir.

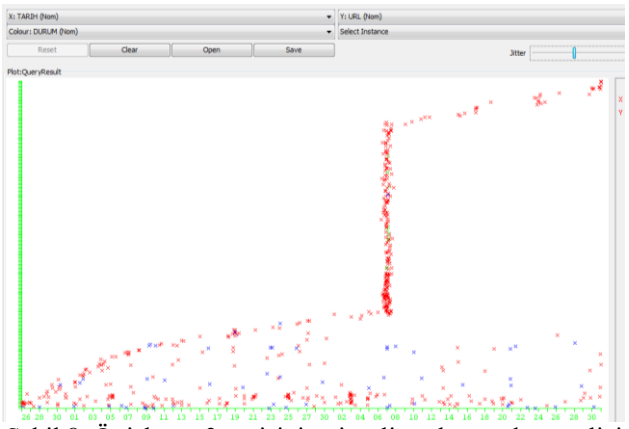
Birliktelik kuralları konusundaki en meşhur algoritmalardan biri Apriori algoritmasıdır [44]. Önişleme-2 verisinde tarih, URL ve durum alanları için en az destek değeri 0.1 ve en az güven değeri 0.9 için Apriori algoritmasıyla başarılı (200 durum kodu) ve en sık ziyaret edilen sayfalar arasında birliktelik kuralları oluştuğu görülmüştür.

4.3. WEKA ile Saldırı Tespiti (Intrusion detection with WEKA)

Bir önceki bölümde WEKA ile istatistiksel analiz yapılırken kümeleme ve birliktelik kurallarına göre herhangi bir saldırı girişimi tespit edilememiştir. Sadece WEKA'nın Visualize sekmesinden saldırı girişiminde bulunulduğu tahmin edilmiştir.

Bu bölümde doğrudan saldırı tespitine yönelik sadece 400 serili durum kodlarının bulunduğu log satırlarını içeren önişleme-3 verisi kullanılarak analiz yapılmıştır. Veri WEKA'ya yüklendikten sonra önişleme sekmesinde, aynı IP'den çok fazla sayıda hatalı istek alındığı görülmüştür.

Şekil 8'de visualize sekmesinde ise tarih, URL ve durum alanları için yayılım dağılımı incelendiğinde aynı tarihte çok farklı URL'lere istek yapıldığı ve çok sayıda hata durum kodu alındığı görülmektedir. Bu alandaki verilerin detaylı bilgileri incelendiğinde saldırı girişiminde bulunulduğu açıkça görülmektedir.



Şekil 8. Önişleme-3 verisinin visualize ekranından analizi
(Analysis of preprocessing-3 data from the visualizer screen)

Önişleme-3 verisine Apriori algoritmasını çalıştırarak en az destek değeri 0.1 ve en az güven değeri 0.9 için birliktelik kurallarını uyguladığımızda Şekil 9'daki sonuç elde edilmiştir.

Buradaki sonuçlara göre; 65.111.180.205 201 numaralı IP numarasının 07 Ekim 2012 tarihinde sayfayı ziyaret etmiş olduğu ve ilk satırda 201 kural oluşurken ikinci satırda durum kodu 404 olan istekler için 192 satır oluştuğu görülmektedir. Yani aynı IP'den gönderilen isteklerin hepsi ilk kez 07 Ekim 2012 tarihinde gönderilmiş ve neredeyse isteklerin tamamında 404 ile sayfa görüntülenememiştir.

```
 1. IPNO=65.111.180.205 201 ==> TARIH=07/Oct/2012 201    conf:(1)
 2. IPNO=65.111.180.205 DURUM=404 192 ==> TARIH=07/Oct/2012 192    conf:(1)
 3. URL="GET /undefined HTTP/1.1" 77 ==> DURUM=404 77    conf:(1)
 4. TARIH=07/Oct/2012 DURUM=404 193 ==> IPNO=65.111.180.205 192    conf:(0.99)
 5. TARIH=07/Oct/2012 207 ==> IPNO=65.111.180.205 201    conf:(0.97)
 6. IPNO=65.111.180.205 201 ==> DURUM=404 192    conf:(0.96)
 7. IPNO=65.111.180.205 TARIH=07/Oct/2012 201 ==> DURUM=404 192    conf:(0.96)
 8. IPNO=65.111.180.205 201 ==> TARIH=07/Oct/2012 DURUM=404 192    conf:(0.96)
 9. TARIH=07/Oct/2012 207 ==> DURUM=404 193    conf:(0.93)
10. TARIH=07/Oct/2012 207 ==> IPNO=65.111.180.205 DURUM=404 192    conf:(0.93)
```

Şekil 9. Apriori algoritması sonucunda oluşturulan kurallar (The rules obtained by Apriori algorithm)

Tarih, IP numarası ve durum alanları için Kmeans kümeleme algoritması çalıştırılmış ve iki küme oluştuğu görülmüştür. Şekil 10'da elde edilen sonuç sunulmuştur. Kümelerden biri daha önce saldırı girişiminde bulunduğu tespit edilen IP numarasını içermektedir. Diğer küme ise çok az sayıda veri içermektedir.

```
Cluster centroids:
                                          Cluster#
Attribute               Full Data               0               1
                          (363)             (360)             (3)
=============================================================================
IPNO                65.111.180.205  65.111.180.205  159.146.9.224
TARIH                 07/Oct/2012     07/Oct/2012     20/Oct/2012
DURUM                         404             404             404
```

Şekil 10. K-means ile elde edilen kümeler (Clusters obtained by K-means)

Bu bölümde WEKA ile saldırı girişiminde bulunduğu kesin olarak tespit edilmiştir. Bu analizlerden elde edilen saldırı girişiminin tespit edildiği tarih, durum kodu, URL gibi bilgiler web logları saldırı analiz aracı olan Apache

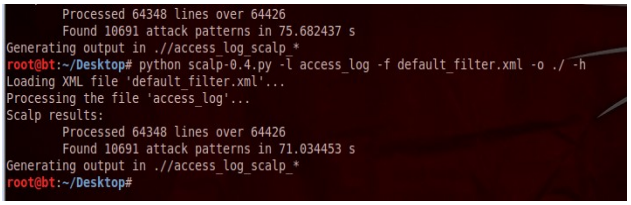
Scalp ile saldırı türünü ve saldırının başarılı olup olmadığını tespit etmek üzere kullanılmıştır.

5. APACHE SCALP İLE LOG ANALİZİ *(LOG ANALYSIS BY USING APACHE SCALP)*

Scalp, güvenlik uzmanlarının kullandığı, Apache log analiz aracıdır. Scalp, web sunucunun güvenliğini logları analiz ederek değerlendiren ve bunu gerçekleştirmek üzere düzenli ifadeler kullanan bir phyton scripttir. Web uygulamalarının SQL Enjeksiyon, XSS, CSRF vb. bilinen saldırıların hedefinde olup olmadığını analiz etmektedir [45]. Bu script PHPIDS düzenli ifadelerini kullanmaktadır; bu ifadeler filtreleme seçeneğini sağlayan "default_filter.xml" dosyasında yer almaktadır [46].

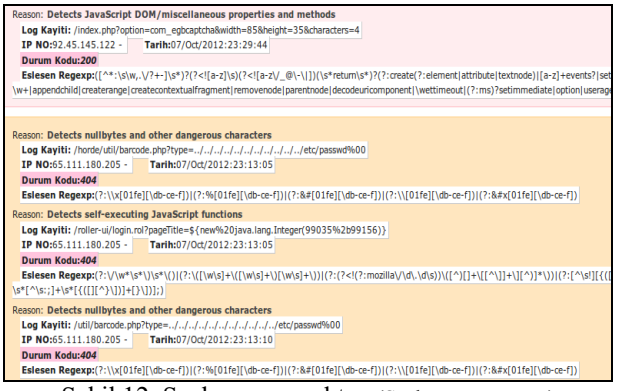
Bu çalışmada Scalp'ın tercih edilme sebebi Scalp'ın doğrudan Apache aracı olup web logları saldırı tespitine yönelik geliştirilmiş olmasından kaynaklanmaktadır.

Bu aşamada Scalp'da önişleme-2 verisi analiz edilecektir. WEKA'dan web madenciliği işlemleri sonucunda elde edilen analiz sonuçları Scalp'da filtreleme yapmaya yardımcı olacaktır. Son işlenen veri olan önişleme-3 verisi yerine önişleme-2 verisinin kullanılma sebebi; saldırı girişimlerinin başarılı olup olmadığını tespit etmek içindir. Çünkü 400 serili durum kodlarının yer aldığı önişleme-3 verisinde saldırı girişimleri tespit edilebilmekte ancak saldırının başarılı olup olmadığını anlamak için 200 serili durum kodlarının yer aldığı işlenmiş olan önişleme-2 verisinin kullanılması gerekmektedir. Önişleme-2 verisini analiz etmek üzere, Backtrack kullanılarak Şekil 11'de gösterilen script çalıştırılmıştır.



Şekil 11. Scalp ile script çalıştırma ve sonuçları (Script execution and results of the scalp)

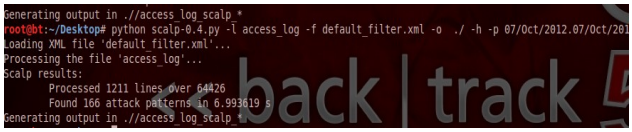
Toplam 64348 satır için işlem yapılmıştır ve 10691 saldırı girişimi 71 saniye içerisinde tespit edilmiştir. Dosyanın içeriğinden bir kesit Şekil 12'de sunulmuştur.



Şekil 12. Scalp raporu çıktısı (Scalp report output)

Scalp ile üretilen sonuçların yaklaşık %15'i "false positive" yani yanlış alarm olmaktadır. Dolayısıyla sonuçları analiz eden kişinin dikkatli bir incelemeden geçirmesi gerekmektedir. Bu ise veri boyutunun çok büyük olduğu durumlarda 100.000 lerce satır arasından saldırı tesptini yapmayı gerektirmektedir. Bu noktada bir önceki bölümde web madenciliği ile analiz edilmiş olan veri sonuçlarına göre filtreleme yapılarak Scalp'da gerçekleştirilecek saldırı analizinde veri boyutu ciddi oranda azaltılıp; analizi çok daha kısa sürede tamamlamak hedeflenmiştir.

WEKA'da gerçekleştirilen örüntü keşfinde hem visualize hem de kümeleme ve birliktelik kurallarıyla saldırı girişimi olduğu tespit edilen isteklerin tarihi 7 Ekim 2012'dir. Scalp'da 7 Ekim 2012 tarihine göre filtreleme yapılarak analiz tekrar yapılmıştır ve Şekil 13'teki sonuçlar elde edilmiştir. Toplam 1211 satır işlenmiştir ve 7 saniye içerisinde 166 saldırı girişimi tespit edilmiştir.



Şekil 13. Tarihe göre filtrelenerek çalıştırılan script (Scripts filtered by date)

Tablo 3'de Scalp analiz sonuçlarının karşılaştırılması yapılmıştır.

Tablo 3. Scalp analiz sonuçlarının karşılaştırılması (Comparison of Scalp Analysis)

	Önişleme-2 Verisi Analiz Sonuçları	Örüntü keşfi sonucuna göre filtrelenen veri
İşlenen Satır Sayısı	64348	1211
Veri İşleme Süresi	71 saniye	7 saniye
Saldırı Girişimi Olabilecek Satır Sayısı	10691	166

İşlem süresi %90.1 oranında kısalırken, incelenecek saldırı girişimleri sayısı da %88.7 oranında azalmıştır. Apache Scalp'da saldırının etki değeri zarar verme derecesine göre 1-10 değerleri arasında derecelendirilmiştir. Apache Scalp ile tespit edilen saldırı türleri ve sayıları etki değerleri ile birlikte Tablo 4'de sunulmuştur.

Tablo 4. Apache Scalp ile tespit edilen saldırı girişimleri
(The attack attempts detected by Apache Scalp)

Saldırı Türleri	Saldırı Sayısı	Etki Değerleri (1-10 arası)
XSS	163	3,5,6
SQL Enjeksiyon	2	5,6
Sunucudan Dosya Çağırma (LFI)	10	5,6

6. SONUÇ VE ÖNERİLER (CONCLUSIONS AND RECOMMENDATIONS)

Gerçekleştirilen istatistiksel analizler neticesinde kullanıcı eğilimleri ve genel site kullanımı hakkında elde edilen bulgularla site geliştiricilerine web sitesinin iyileştirilmesine ve geliştirilmesine yönelik katkıda bulunulacağı düşünülmektedir. WEKA'da gerçekleştirilen analizlerde hem web robotlarıyla hem de web robotları olmadan analiz gerçekleştirilerek ziyaretçi sayısı, başarılı istek sayısı, hatalı istek sayısı gibi sonuçlar elde edilmiştir. Kümeleme ve birliktelik kuralları ile belirli tarihlerde durum kodlarına göre en çok ziyaret edilen sayfalar belirlenmiştir.

Elde edilen sonuçlar ve öneriler aşağıda sunulmuştur:

- Genel olarak sayfalara yapılan ziyaretler, sayfa görüntüleme ve bant genişliği hakkında istatistiksel bilgiler verilmiştir. Sayfaya yapılan ziyaretlerde web robotlarının gönderdiği isteklerle kullanıcı isteklerinin birbirine çok yakın sayıda olduğu görülmüştür. Özellikle kullanıcılarla ilgili yapılacak analizlerde yanıltıcı olmaması için web robotlarının kayıtlarının temizlenmesi gerektiği sonucuna varılmıştır.
- Web robotları sayfaya kısa süre içerisinde çok fazla ve arka arkaya istek gönderdiğinden web robotlarının da yer aldığı önişleme verisi için analiz gerçekleştirildiğinde hata serili durum kodlarıyla visualize ekranında kümeler oluştuğu görülmüştür. Toplam 6 kümeden 5'i web robotlarına ait iken sadece 1 kümenin gerçek saldırı girişimi olduğu tespit edilmiştir. İkinci önişleme verisinden web robotları çıkarıldıktan sonra sadece tek bir küme oluştuğu görülmüştür. Özellikle web madenciliği ile saldırı analizi yapılacağında web robotlarının önişleme aşamasında belirlenip indekslediği sayfaların temizlenmesi gerektiği sonucuna varılmıştır.
- Web robotlarının temizlenmiş olduğu önişleme-2 verisi web madenciliği ile analiz edildiğinde 200 durum koduyla en çok ziyaret edilen sayfalar tespit edilmiştir. Web loglarında geniş tarih aralığında işlem yapıldığında çoğu zaman işlemin başarılı olduğuna dair durum kodu alınacağından visualize ekranında oluşan saldırı örüntüleri kümeleme veya birliktelik kuralları ile görülememiştir. Bu nedenle web madencilliğinde özellikle saldırı tespitine yönelik analiz gerçekleştirileceğinde 400 serili hata durum kodlarının haricindeki log satırlarının temizlenmesi gerekmektedir. Bu şekilde elde edilen önişlem-3 verisine web madenciliği algoritmaları uygulandığında doğrudan saldırı tespiti yapılmıştır.
- Web madenciliği ile elde edilen saldırı örüntüleri ile ilgili bilgiler, Apache Scalp aracında filtreleme yapmak üzere kullanılmış ve saldırıların türü ve yeri ile ilgili detaylı analizler yapılmıştır. İşlem süresi %90.1 oranında kısalırken, incelenecek saldırı girişimleri sayısı da %88.7 oranında azalmıştır.
- Scalp'ın sonuçlarına göre en çok XSS ve SQLI atakları yapıldığı görülmüştür. Saldırganlar gizlenebilmek amacıyla çeşitli encoding yöntemlerini kullanarak aranacak kelime gruplarının farklı şekillerde log dosyasında tutulmasını sağlayabilmektedir. Site geliştiricilerin bu atakların bertaraf edilmesine yönelik kaynak kodunda gerekli kontrolleri yapmaları ve varsa saldırı tespit sisteminde gerekli konfigürasyonları uygulamaları gerekmektedir. Saldırı girişimleri daha çok bilinen açıklıklar üzerinden gerçekleştirilmektedir, dolayısıyla yazılımların güncel tutulması, yamaların yapılması da oldukça önemlidir.

Bir sonraki çalışamada, web loglarından saldırı tespit işleminin tam otomatize gerçekleştirilmesi düşünülmektedir.

KAYNAKLAR (REFERENCES)

[1] A. Adamov, "Data mining and analysis in depth. case study of Qafqaz University HTTP server log analysis", ***In Application of Information and Communication Technologies (AICT) IEEE 8th International Conference***, 1-4, 2014.

[2] A. S. Yumnam, Y. Chaitanya Sreeram & S. A. Naeem, "Overview: Weblog mining, privacy issues and application of Web Log mining", ***In Computing for Sustainable Global Development (INDIACom) International Conference***, 638-641, 2014.

[3] A. D. Khairkar, D. D. Kshirsagar & S. Kumar, "Ontology for Detection of Web Attacks", **In Communication Systems and Network Technologies (CSNT) International Conference**, 612-615, 2013.

[4] B. Mobasher, R. Cooley, J. Srivastava, "Automatic Personalization based on Web Usage Mining", Communications of the ACM, 43(8),142-151, 2000.

[5] A. Vahaplar, M. M. İnceoğlu, "Veri Madenciliği ve Elektronik Ticaret", **Türkiye'de Internet Konferansları VII**, 2001.

[6] R. Daş, **Web Kullanıcı Erişim Kütüklerinden Bilgi Çıkarımı**, Doktora Tezi, Fırat Üniversitesi Fen Bilimleri Enstitüsü, Elazığ, 2008.

[7] J. P. Leite, "Analysis of log files as a security aid", **In Information Systems and Technologies (CISTI) IEEE 6th Iberian Conference**, 1-6, 2011.

[8] M. Nagappan & M. A. Vouk, "Abstracting log lines to log event types for mining software system logs", **In Mining Software Repositories 7th IEEE Working Conference**, 114-117, 2010.

[9] K.C. Burçak, **Kırıkkale Üniversitesi Web Sitesinin Kullanıcı Örüntülerinin Web Madenciliği ile Analizi**, Yüksek Lisans Tezi, Kırıkkale Üniversitesi Fen Bilimleri Enstitüsü, Kırıkkale, 2012.

[10] S. S. Vernekar & A. Buchade, "MapReduce based log file analysis for system threats and problem identification", **In Advance Computing Conference (IACC)**, 831-835, 2013.

[11] O. Ozulku, N. F. Fadhel, D. Argles & G. B. Wills, "Anomaly detection system: Towards a framework for enterprise log management of security services", **In Internet Security (WorldCIS) 2014 World Congress**, 97-102, 2014.

[12] İ. Haberal, **Veri Madenciliği Algoritmaları Kullanılarak Web Günlük Erişimlerinin Analizi**, Yüksek Lisans Tezi, Başkent Üniversitesi Fen Bilimleri Enstitüsü, Ankara, 2007.

[13] T. Hussain, S. Asghar & N. Masood, "Web usage mining: A survey on preprocessing of web log file", **In International Conference on Information and Emerging Technologies (ICIET)**, 1-6, 2010.

[14] B. Özakar & H. Püskülcü, "Web içerik ve web kullanım madenciliği tekniklerinin entegrasyonu ile oluşmuş bir veri tabanından nasıl yararlanılabilir?", 2002.

[15] D. S. Sisodia & S. Verma, "Web usage pattern analysis through web logs: A review", **In Computer Science and Software Engineering (JCSSE) International Joint Conference**, 49-53, 2012.

[16] R. Yevale, M. Dhage, T. Nalawade & T. Kaule, "Unauthorized Terror Attack Tracking Using Web Usage Mining", (IJCSIT) International Journal of Computer Science and Information Technologies, 5 (2), 1210-1212, 2014.

[17] T. Gržinic, , T. Kišasondi, J. Šaban, "Detecting anomalous Web server usage through mining access logs", **Central European Conference on Information and Intelligent Systems**, 228-296, 2013.

[18] M. Turan, **Web Mining: Pattern Discovery On The World Wide Web,** Yüksek Lisans Tezi, Dokuz Eylül Üniversitesi Fen Bilimleri Enstitüsü, İzmir, 2011.

[19] G. Sarıman, **Paralel Programlama ile Web Madenciliğinde Log Analizi**, Yüksek Lisans Tezi, Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü, Isparta, 2011.

[20] İnternet: H. Önal, Web Sunucu Loglarından Saldırı Analizi, http://blog.bga.com.tr/2013/01/web-sunucu-loglarndan-saldr-analizi.html, 20.11.2014.

[21] G. Vigna, W. Robertson, V. Kher & R. A. Kemmerer, "A stateful intrusion detection system for world-wide web servers", **In Computer Security Applications Conference**, 34-43, 2003.

[22] M. Auxilia & D. Tamilselvan, "Anomaly detection using negative security model in web application", **In Computer Information Systems and Industrial Management Applications (CISIM) International Conference**, 481-486, 2010.

[23] Z. Sun, H.Sheng, M. Wei, J. Yang, H. Zhang & L. Wang, "Application of web log mining in local area network security", **In Electronic and Mechanical Engineering and Information Technology (EMEIT) International Conference**, 8, 3897-3900, 2011.

[24] E. Yıldız, **Veri Madenciliği Teknikleriyle Saldırı Tespiti ve Bir Uygulama**, Yüksek Lisans Tezi, Gazi Üniversitesi Bilişim Enstitüsü, Ankara, 2010.

[25] S. E. Salama, M. I. Marie, L. M.El-Fangary & Y. K. Helmy, "Web Server Logs Preprocessing for Web Intrusion Detection", *Computer and Information Science*, 4(4), 123, 2011.

[26] P. V. Patil & D. R. Patil, "Preprocessing Web Logs For Web Intrusion Detection", *International Journal of Applied Information Systems (IJAIS)*, 2012.

[27] M. Mabzool, M. Z. Lighvan, Intrusion Detection System Based On Web Usage Mining, *International Journal of Computer Science*, 4(1), 2014.

[28] İnternet: OWASP Top Ten Project, https://www.owasp.org/index.php/Category:OWASP_Top_Ten_Project, 08.12.2014.

[29] İnternet: OWASP Top Ten Project, https://www.owasp.org/index.php/Top_10_2013-Top_10, 09.12.2014.

[30] İnternet: Common Vulnerabilities and Exposures (The Standard for Information Security Vulnerability Names). http://cwe.mitre.org/, 08.12.2014.

[31] R. Johari, P.Sharma, "A survey on web application vulnerabilities (SQLIA, XSS) exploitation and security engine for SQL Enjeksiyon", **Communication Systems and Network Technologies (CSNT) International Conference**, 453-458, 2012.

[32] İnternet: Infosec Institute. OWASP Top 10 Tools and Tactics. http://resources.infosecinstitute.com/owasp-top-10-tools-and-tactics/, 08.12.2014

[33] P. Hernandez, I. Garrigos & , J. Mazón, "Modeling web logs to enhance the analysis of Web usage data", **In Database and Expert Systems Applications (DEXA) IEEE 2010 Workshop**, 297-301, 2010.

[34] A. Guerbas, O. Addam, O. Zaarour, M. Nagi, A. Elhajj, M. Ridley, R. Alhajj, "Effective web log mining and online navigational pattern prediction", Knowledge-Based Systems, 49, 50-62, 2013.

[35] H. Arslan, **Sakarya Üniversitesi Web Sitesi Erişim Kayıtlarının Web Madenciliği ile Analizi**, Yüksek Lisans Tezi, Sakarya Üniversitesi Fen Bilimleri Enstitüsü, Sakarya, 2008.

[36] T. T. Aye, "Web log cleaning for mining of web usage patterns", **In Computer Research and Development (ICCRD) IEEE 2011 3rd International Conference**, 2, 490-494, 2011.

[37] M. Shu-yue, L. Wen-cai & , W. Shuo, "The Study on the Preprocessing in Web Log Mining", **In Knowledge Acquisition and Modeling (KAM) IEEE 2011 Fourth International Symposium**, 315-317, 2011.

[38] M. P. Yadav, P. K. Keserwani & S. G. Samaddar, "An efficient web mining algorithm for Web Log analysis: E-Web Miner", **In Recent Advances in Information Technology (RAIT) 2012 1st International Conference**, 607-613, 2012.

[39] A. Al-Hamamı, A. Mohammad, S. Hassan, "Applying data mining techniques in intrusion detection system on web and analysis of web usage", *Information Technology Journal*, 5(1), 1-4, 2006.

[40] C. I. Ezeife, J. Dong, & A. K. Aggarwal, "SensorWebIDS: a web mining intrusion detection system", *International Journal of Web Information Systems*, 4(1), 97-120, 2008.

[41] İnternet: WEKA, http://weka.wikispaces.com/Primer, 06.01.2012.

[42] M. Dener, M. Dörterler, A. Orman, "Açık Kaynak Kodlu Veri Madenciliği Programları: Weka'da Örnek Uygulama", Akademik Bilişim, 9, 11-13, 2009.

[43] İnternet: http://research.cs.queensu.ca/home/cisc333/tutorial/Weka.html, 03.04.2015.

[44] M. Karabatak, M. C. İnce, "Apriori Algoritması ile Öğrenci Başarısı Analizi", *Elektrik Elektronik Bilgisayar Mühendisliği Sempozyumu*, 8-12, 2004.

[45] İnternet: Apache Log Analyser for Security Specialists, http://www.infosec.lk/2011/02/scalp-apache-log-analyser-for-security-specialists.html, 24.08.2014.

[46] İnternet: Apache Log Analyser for Security apache-scalp,http://code.google.com/p/apache-scalp/source/browse/trunk/default_filter.xml, 24.08.2014.

[47] T. Ozseven,M.Düğenci, " LOG PreProcessing: Web Kullanım Madenciliği Ön İşlem Aşaması Uygulama Yazılımı", Gazi Üniversitesi Bilişim Teknolojileri Dergisi, 4(2):55-66, 2011,

[48] R. Daş., İ. Türkoğlu, "Creating meaningful data from web logs for improving the impressiveness of a website by using path analysis method", Expert Systems with Applications, 36(3), 6635-6644, 2009.

[49] S. Araya, , M. Silva ve R. Weber, "A methodology for web usage mining and its application to target group identification", Fuzzy sets and systems, 148(1),139-152, 2004.

[50] R. Meyer, "Detecting Attacks on Web Applications from Log Files" SANS Institute InfoSec Reading Room, 2008.

[51] W. Win, H. H. Htun, "A Simple and Efficient Framework for Detection of SQL Enjeksiyon Attack". IJCCER, 1(2), 26-30, 2013.

[52] Y. M. Mali, R. M. JV, M. Raj, A. T. Gaykar, "Honeypot: a tool to track hackers", **IRACST – Engineering Science and Technology: An International Journal**,(4), 2250-3498, 2014.

[53] I. Hydara, A. B. M. Sultan, H. Zulzalil, N. Admodisastro, "Current state of research on cross-site scripting (XSS)–A systematic literature review", Information and Software Technology, 58, 170-186, 2015.

[54] R. Daş İ. Türkoğlu, & M. Poyraz, Genetik Algoritma Yöntemiyle Internet Erişim Kayıtlarından Bilgi Çıkarılması. Sakarya Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 10(2), 67-72, 2006.